Volume II, Appendix C
Table of Contents

C

Appendix C: Qualification Test Design Criteria

C.1 Scope

This appendix describes the guiding principles used to design the voting system qualification testing process conducted by ITAs.

Qualification tests are designed to demonstrate that the system meets or exceeds the requirements of the Standards. The tests are also used to demonstrate compliance with other levels of performance claimed by the manufacturer.

Qualification tests must satisfy two separate and possibly conflicting sets of considerations. The first is the need to produce enough test data to provide confidence in the validity of the test and its apparent outcome. The second is the need to achieve a meaningful test at a reasonable cost, and cost varies with the difficulty of simulating expected real-world operating conditions and with test duration. It is the test designer's job to achieve an acceptable balance of these constraints.

The rationale and statistical methods of the test designs contained in the Standards are discussed below. Technical descriptions of their design can be found in any of several books on testing and statistical analysis.

C.2 Approach to Test Design

The qualification tests specified in the Standards are primarily concerned with assessing the magnitude of random errors. They are also, however, capable of detecting bias errors that would result in the rejection of the system.

Test data typically produce two results. The first is an estimate of the true value of some system attribute such as speed, error rate, etc. The second is the degree of certainty that the estimate is a correct one. The estimate of an attribute's value may or may not be greatly affected by the duration of the test. Test duration, however, is very important to the degree of certainty; as the length of the test increases, the level of uncertainty decreases. An efficient test design will produce enough data over a sufficient period of time to enable an estimate at the desired level of confidence.

There are several ways to design tests. One approach involves the preselection of some test parameter, such as the number of failures or other detectable factor. The essential element of this type of design is that the number of observations is independent of their results. The test may be designed to terminate after 1,000 hours or 10 days, or when 5 failures have been observed. The number of failures is important because the confidence interval (uncertainty band) decreases rapidly as the number of failures increases. However, if the system is highly reliable or very accurate, the length of time required to produce a predetermined number of failures or errors using this method may be unachievably long.

Another approach is to determine that the actual value of some attribute need not be learned by testing, provided that the value can be shown to be better than some level. The test would not be designed to produce an estimate of the true value of the attribute but instead to show, for example, that reliability is at least 123 hours or the error rate is no greater than one in ten million characters.

The latter design approach, which was chosen for the Standards, uses what is called Sequential Analysis. Instead of the test duration being fixed, it varies depending on the outcome of a series of observations. The test is terminated as soon as a statistically valid decision can be reached that the factor being tested is at least as good as or no worse than the predetermined target value. A sequential analysis test design called the "Wald Probability Ratio Test" is used for reliability and accuracy testing.

C.3 Probability Ratio Sequential Test (PRST)

The design of a Probability Ratio Sequential Test (PRST) requires that four parameters be specified:

H0, the null hypothesis
H1, the alternate hypothesis

a, the Producer's risk
b, the Consumer's risk

The Standards anticipate using the PRST for testing both time-based and event-based failures.

This test design provides decision criteria for accepting or rejecting one of two test hypotheses: the null hypothesis, which is the Nominal Specification Value (NSV), or the alternate hypothesis, which is the MAV. The MAV could be either the Minimum Acceptable Value or the Maximum Acceptable Value depending upon what is being tested. (Performance may be specified by means of a single value or by two values. When a single value is specified, it shall be interpreted as an upper or lower single-sided 90 percent confidence limit. If two values, these shall be interpreted as a two-sided 90 percent confidence interval, consisting of the NSV and MAV.)

In the case of Mean Time Between Failure (MTBF), for example, the null hypothesis is that the true MTBF is at least as great as the desired value (NSV), while the alternate hypothesis is that the true value of the MTBF is less than some lower value (Minimum Acceptable Value). In the case of error rate, the null hypothesis is that the true error rate is less than some very small desired value (NSV), while the alternate hypothesis is that the true error rate is greater than some larger value that is the upper limit for acceptable error (Maximum Acceptable Value).

C.4 Time-based Failure Testing Criteria

An equivalence between a number of events and a time period can be established when the operating scenarios of a system can be determined with precision. Some of the performance test criteria of Volume II, Section 4, *Hardware Testing*, use this equivalence.

System acceptance or rejection can be determined by observing the number of relevant failures that occur during equipment operation. The probability ratio for this test is derived from the Exponential probability distribution. This distribution implies a constant hazard rate for equipment failure that is not dependent on the time of testing or the previous failures. In that case, two or more systems may be tested simultaneously to accumulate the required number of test hours, and the validity of the data is not affected by the number of operating hours on a particular unit of equipment. However, for environmental operating hardware tests, no unit shall be subjected to less than two complete 24 hour test cycles in a test chamber as required by Volume II, Subsection 4.7.1 of the Standards.

In this case, the null hypothesis is that the Mean Time Between Failure (MTBF), as defined in Volume I, Subsection 3.4.3 of the Standards, is at least as great as some value, here the Nominal Specification Value. The alternate hypothesis is that the MTBF is no better than some value, here the Minimum Acceptable Value.

For example, a typical system operations scenario for environmental operating hardware tests will consist of approximately 45 hours of equipment operation. Broken down, this time allotment involves 30 hours of equipment set-up and readiness testing and 15 hours of elections operations. If the Minimum Acceptable Value is defined as 45 hours, and a test discrimination ratio of 3 is used (in order to produce an acceptably short expected time of decision), then the Nominal Specification Value equals 135 hours.

With a value of decision risk equal to 10 percent, there is no more than a 10 percent chance that a system would be rejected when, in fact, with a true MTBF of at least 135 hours, the system would be acceptable. It also means that there is no more than a 10 percent chance that a system would be accepted with a true MTBF lower than 45 hours when it should have been rejected.

Therefore,

H0: MTBF = 135 hours
H1: MTBF = 45 hours

a = 0.10
b = 0.10.

Under this PRST design, the test is terminated and an ACCEPT decision is reached when the cumulative number of equipment hours in the second column of the following table has been reached, and the number of failures is equal to or less than the number shown in the first column. The test is terminated and a REJECT decision is reached when the number of failures occurs in less than the number of hours specified in the third column. Here, the minimum time to accept (on zero failures) is 169 hours. In the event that no decision has been reached by the times shown in the last table entries, the test is terminated, and the decision is declared as indicated. Any time that 7 or more failures occur, the test is terminated and the equipment rejected. If after 466 hours of operation the cumulative failure score is less than 7.0, then the equipment is accepted.

Number of Failures	Accept if Time Greater Than	Reject if Time Less Than
0	169	Continue test
1	243	Continue test
2	317	26
3	392	100
4	466	175
5	466	249
6	466	323
7	N/A	(1)

(1) Terminate and REJECT

This test is based on the table of test times of the truncated PRST design V-D in the Military Handbook MIL-HDBK-781A that is designated for discrimination ratio 3 and a nominal value of 0.10 for both a and b. The Handbook states that the true producer risk is 0.111 and the true consumer risk is 0.109. Using the theoretical formulas for either the untruncated or truncated tests will lead to different numbers.

The test design will change if given a different set of parameters. Some jurisdictions may find the Minimum Acceptable Value of 45 hours unacceptable for their needs. In addition, it may be appropriate to use a different discrimination ratio, or different Consumer’s and Producer’s risk. Also, before using tests based on the MTBF, it should be determined whether time-based testing is appropriate rather than event-based or another form of testing. If MTBF-based procedures are chosen, then the appropriateness of the assumption of a constant hazard rate with exponential failures should in turn be assessed.

C.5 Accuracy Testing Criteria

Some voting system performance attributes are tested by inducing an event or series of events, and the relative or absolute time intervals between repetitions of the event has no significance. Although an equivalence between a number of events and a time period can be established when the operating scenarios of a system can be determined with precision, another type of test is required when such equivalence cannot be established. It uses event-based failure frequencies to arrive at ACCEPT/REJECT criteria. This test may be performed simultaneously with time-based tests.

For example, the failure of a device is usually dependent on the processing volume that it is required to perform. The elapsed time over which a certain number of actuation cycles occur is, under most circumstances, not important. Another example of such an attribute is the frequency of errors in reading, recording, and processing vote data.

The error frequency, called “ballot position error rate,” applies to such functions as process of detecting the presence or absence of a voting punch or mark, or to the closure of a switch corresponding to the selection of a candidate.

Qualification and acceptance test procedures that accommodate event-based failures are, therefore, based on a discrete, rather than a continuous probability distribution. A Probability Ratio Sequential Test using the binomial distribution is recommended. In the case of ballot position error rate, the calculation for a specific device (and the processing function that relies on that device) is based on:

HO: Desired error rate = 1 in 10,000,000

H1: Maximum acceptable error rate = 1 in 500,000

a = 0.05

b= 0.05

and the minimum error-free sample size to accept for qualification tests is 1,549,703 votes.

The nature of the problem may be illustrated by the following example, using the criteria contained in the Standards for system error rate. A target for the desired accuracy is established at a very low error rate. A threshold for the worst error rate that can be accepted is then fixed at a somewhat higher error rate. Next, the decision risk is chosen, that is the risk that the test results may not be a true indicator of either the system's acceptability or unacceptability. The process is as follows:

- ♦ The desired accuracy of the voting system, whatever its true error rate (which may be far better), is established as no more than one error in every ten million characters (including the null character).
- ♦ If it can be shown that the system's true error rate does not exceed one in every five hundred thousand votes counted, it will be considered acceptable. (This is more than accurate enough to declare the winner correctly in almost every election.)
- ♦ A decision risk of 5 percent is chosen, to be 95 percent sure that the test data will not indicate that the system is bad when it is good or good when it is bad.

This results in the following decision criteria:

- ♦ If the system makes one error before counting 26,997 consecutive ballot positions correctly, it will be rejected. The vendor is then required to improve the system;
- ♦ If the system reads at least 1,549,703 consecutive ballot positions correctly, it will be accepted; and

If the system correctly reads more than 26,997 ballot positions but less than 1,549,703 when the first error occurs, the testing will have to be continued until another 1,576,701 consecutive ballot positions are counted without error (a total of 3,126,404 with one error).